Title	Reliability of professional and naï ve listeners on perceptual evaluation of different types of voice sample in children
Author(s)	Ma, Kam-wa; 馬錦華
Citation	Ma, K. [馬錦華]. (2014). Reliability of professional and naï ve listeners on perceptual evaluation of different types of voice sample in children. (Thesis). University of Hong Kong, Pokfulam, Hong Kong SAR.
Issued Date	2014
URL	http://hdl.handle.net/10722/238938
Rights	This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.; The author retains all proprietary rights, (such as patent rights) and the right to use in future works.

Reliability of Professional and Naïve Listeners on Perceptual Evaluation of Different Types of Voice Sample in Children

Ma Kam Wa

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, June 30, 2014.

Abstract

This study investigated the effects of speech contexts, severity of voice problems and professional background on reliability of perceptual voice evaluation in children. Two groups of listener were recruited. The first group comprised of 10 speech therapists with 1 to 13 years of experience in working with pediatric speech and language caseloads (Professional Group). The second group comprised of 20 naïve listeners (Naïve Group). Both groups of listeners were asked to rate perceptually the severity levels of voice samples of 40 children speakers on three vocal parameters (roughness, breathiness and overall severity). For each child, there were three types of voice samples including prolongation of vowel /a/, sentence and short passage. Intraclass correlation (ICC) was calculated to identify intra- and inter-rater reliability of perceptual voice evaluation by the two groups of listener. Results revealed higher rater reliabilities 1) in passage reading, and 2) by professional listeners. Interestingly, disagreements were noted between the two groups on reliability across voices with different severity levels. These findings contradicted the initial hypothesis of higher reliability on normal and severely dysphonic voices in children.

Keywords: reliability, children, speech context, vocal parameter, severity, professional listeners, naïve listeners

Introduction

Voice problems are evaluated through case history taking, perceptual rating of voice, measurements of physiological characteristics on acoustic, aerodynamic, vibratory and muscle action, and examinations on laryngeal anatomy and physiology. Among the above, determining the severity of perceptual signs of voice problems is one of the major components in initial diagnosis (Colton, Casper, & Leonard, 2006). Apart from that, perceptual rating also serves as a means to document treatment outcomes in voice management. It helps clinicians decide if treatment approaches are suitable by making perceptual rating before and after a treatment session, and helps make decisions throughout the course of continuous management (Kent, 1996; Oates, 2009).

Although perceptual voice rating is generally more cost-effective and convenient than instrumental measurements, it is considered to be subjective. It should be noted that perceptual rating can be affected by a number of factors including 1) environmental factors (e.g., the presence of background noise); 2) personal factors (e.g., listeners' experience, personal preference, cultural and language background, and a shift in internal standard); and 3) study design (e.g., types of rating scale used, speakers' information provided, attributes in voice patterns, scale resolution and the magnitude of target characteristic in the speech sample) (Colton et al., 2006; Freitas, Pestana, Almeida, & Ferreira, 2013; Ghio, Revis, Merienne, & Giovanni, 2013; Kent, 1996; Kreiman, Gerratt, & Ito, 2007; Oates, 2009; Yiu, Murdoch, Hird, Lau, & Ho, 2008).

Types of voice sample investigated in the literatures include sustained vowel prolongation, which was rated on its onset, central part, stable part or as a whole, connected speech which was elicited by sentence and passage reading, and spontaneous speech, including counting from 1 to 10 and spontaneous speech production based on a topic (Baker et al., 2008; de Krom, 1994; Law et al., 2012; Munoz et al., 2002).

There are several factors affecting the quality of voice samples, and they vary under different speech contexts. In sustained vowel prolongation, the voice quality would not be affected by speaker's articulation and changes in speech melody. Production of sustained vowels requires minimal variations in laryngeal and suprasegmental muscles configuration, which provide relatively constant airflow through glottis in phonation. These suggest that sustained vowels are easy to be elicited, but they may not reveal the most natural voice used in daily situations.

On the contrary, the voice quality of connected speech, as in reading and spontaneous speech production, involves coarticulation. Coarticulation requires good coordination of vocal folds, laryngeal and supralaryngeal muscles. Therefore, it is expected that connected speech helps reveal more deviant voice features of a speaker than sustained vowel prolongation (Poletto, Verdum, Strominger, & Ludlow, 2004). Among the two examples of connected speech, spontaneous speech has long been the most natural and representative voice sample because it can better reflect everyday voice of a speaker. However, possible variations owing to situational, emotional or attitudinal factors, such as utterance length, speech melody and the content of production, would increase the difficulties in standardizing and eliciting spontaneous speech sample (Monoz et al., 2002; Swerts, & Veldhuis, 2001). Therefore, it was not commonly used in the studies of voice.

Reliability and agreement are common terms encountered in the studies of perceptual evaluations. According to Kreiman, Gerratt, Kempster, Erman, & Berke (1993), reliability was different from agreement. Reliability is defined as how constant the relationship of two rated voice is, that is, higher reliability requires more parallel ratings to voice samples. However, agreement is commonly defined as whether the ratings are agreed exactly or within a variation of one scale value, that is, high agreement requires holding the same meanings to each scale point on the rating scale.

Regarding whether speech contexts affect rater reliability, studies showed conflicting results on adult population. In an early study, de Krom (1994) examined inter- and intra-rater reliability in four types of voice samples, namely the onset of sustained vowel, sustained vowel without the onset, the entire sustained vowel prolongation and reading. Intra-rater reliability was fairly high across voice samples, without significant effects of the types on perceptual evaluation of voice. Inter-rater reliability was also high across voice samples. These suggested no effect on intra- and inter-rater reliability by speech types.

Munoz et al. (2002) explored the reliability on voices of adult speakers from age 20 to 45. Vowel context was analyzed using its central portion, and was compared with a fragment of connected speech. Results showed good reliability on both vowel and connected speech by professional raters, with a higher reliability on connected speech (i.e. inter-rater reliability). However, intra-rater reliability was fluctuated, implying that the parameters considered by a listener in evaluating vowels were not exactly the same as that in evaluating connected speeches, and thus, speech contexts might affect the reliability on perceptual evaluation of voice.

A recent study by Law et al. (2012), involving adult speakers between aged 24 and 64, concluded with statistically significant differences in intra-rater reliability across sustained vowel, passage reading and conversational speech, with higher intra-rater reliability obtained from connected speech than from sustained vowel. Inter-rater reliability showed statistically insignificant difference across the three types of voice samples, but statistically significant difference between normal and severely dysphonic voice with the highest reliability on severe grade of voice problem. Therefore, speech types and dysphonic severity might influence the reliability on perceptual rating of voice.

The above-mentioned studies focused only on adult speakers, but not on children speakers (de Krom, 1994; Law et al., 2012; Munoz et al. 2002). There is a lack of studies

investigating the reliability of perceptual rating of voice in children. Whether the findings in adult population can be generalized to children needs to be validated.

Regarding whether there are differences between reliability of perceptual rating of voice by different groups of listener, i.e. professional listeners, naïve listeners and dysphonic individuals, studies revealed contradictory results. Karnell et al. (2007) found that the reliability of professional ratings on overall severity of dysphonia was higher and was less affected by the types of rating tool, but was vise versa for individuals with dysphonia. On the contrary, Eadie et al. (2010) concluded that there was no statistical difference on reliability across groups (experienced listeners, inexperienced listeners and dysphonic individuals) for perceptual rating on overall severity. There were no significant relationships found between any demographic factors, such as number of years of voice-related clinical experience, and listeners' judgments, but there were strong relationships between judgments made by experienced and inexperienced listeners and weak-to-moderate relationships between judgments made by dysphonic individuals and other listeners.

In summary, there are conflicting conclusions drawn in the literatures regarding the effect of speech contexts and the listeners' background on the reliability of perceptual ratings of adult's voice qualities. Moreover, there is a lack of studies on the reliability of perceptual evaluation of voice problems in children. Therefore, the present study aims to examine inter- and intra-rater reliability on children's voice quality by listeners with different backgrounds (i.e., professional and naïve listeners) on different speech contexts (i.e., sustained vowel, sentence reading and passage reading). Speakers with different grades of dysphonic severity (i.e., normal, mild, moderate and severe) are included in this study. It is hypothesized that higher rater reliability would be obtained by professional listeners, with higher reliabilities on passage reading by both groups of listener.

Methods

Listeners

Two groups of listener participated in this study. Table 1 lists the background information of professional listeners. The first group (Professional Listener Group) included 10 speech therapists (2 males and 8 females, aged from 23 to 34) with a range of 1 to 13 years of experience in working with pediatric speech and language caseloads (mean = 5.8 years, SD = 4.59 years). All of them self-reported to have normal hearing.

The second group (Naïve Listener Group) included 20 listeners (4 males and 16 females, aged from 20 to 24) with no training on speech and hearing sciences, or linguistics, and with no previous exposure to pathological voices. The ratio of female to male listeners was the same as that in professional listeners. Their hearing abilities were first screened by an audiometer (GSI 18, Grason-Stadler, USA). 75% of them passed the hearing screening at 30 dB for the octave frequencies of 250 Hz to 8000 Hz. For the remaining naïve listeners, the highest intensity level for minimal hearing was 50 dB.

Table 1. Demographic and Clinical Experiences of the Ten Professional Listeners.

Listeners	Sex	Age	Years of Practicing	Percentage of pediatric Voice Cases in current caseload
P1	F	25	3	10%
P2	F	28	6	3%
P3	F	35	13	20%
P4	F	27	4.5	10%
P5	F	29	6.5	0%
P6	F	24	1	20%
P7	F	24	1	0%
P8	F	23	1	0%
P9	M	34	10	10%
P10	M	34	12	2%
			Mean (SD) = 5.8 years (4.59 years)	Mean (SD) = 7.5% (7.82%)

Note: P, professional listeners; M, male; F, female

Speakers and Voice Samples

Samples were drawn from an existing pediatric voice database at the Voice Research Laboratory of the University of Hong Kong. A total of 40 children speakers (15 boys and 25 girls, aged from 6;02 to 12;09) with 16 normal voice speakers and 24 dysphonic speakers were selected. Speakers with speech errors and accents were excluded. The dysphonic speakers were then separated into three severity levels of mild, moderate and severe dysphonia (Nine mildly dysphonic, nine moderately dysphonic and six severely dysphonic). As listeners usually agree more on the constituents in normal or severely dysphonic voices, but disagree more on the extent of mild-to-moderate vocal behaviors (Kreiman, et al., 1993), a spectrum of voice samples, i.e. from normal to severe dysphonic samples, was included to normalize the effect of dysphonic severity on reliability and to simulate daily situations with different severity levels of dysphonia.

Each of the speakers produced three types of voice samples: 1) sustained vowel /a/; 2) sentence: a five-syllable sentence; and 3) passage: the first two complete sentences from a passage. This yielded a total of 120 voice samples. To assess intra-rater reliability, 50% of the voice samples were selected randomly for repetition according to the proportion of speakers in different severity levels, i.e. eight from normal speakers; four from mild, five from moderate and three from severe dysphonic speakers, to give a final total of 180 voice samples. The voice stimuli were all standardized on sound intensity using Audacity 2.0.5 for Mac. Sustained /a/ and sentences were also treated by adding 0.5 seconds of silence before and after the voice.

Rating Scale and Procedures

Listeners were asked to rate the severity on three vocal parameters (roughness, breathiness and overall severity). In this study, roughness was defined as a perception of irregular and

uneven voice quality with the presence of a low frequency noise component and a lack of clarity; breathiness was defined as a perception of glottal air leakage and audible turbulence during phonation; and overall severity was defined as the overall degree of voice abnormality determined (de Krom, 1994; Law, et al., 2012).

An 11-point equal-appearing interval (EAI) scale from 0 (normal) to 10 (severe) was used. It was because Yiu and Ng (2004) suggested that breathiness and roughness could be revealed similarly using EAI and visual analogue (VA) scales. EAI scale also provides higher degrees of convenience and greater ease in using it. In addition, de Krom (1994) concluded that the use of a 4 or 5-point interval scales might be too coarse, scales with a higher resolution, such as a 10 or 11-point scales, might reflect the perceptual ratios more holistically. Therefore, the scales used in this study were 11-point EAI scales, ranging from 0 to 10.

Listeners were provided with the age and gender of the speakers. They were presented with different orders of speech context, with one-third of the listeners started with sustained vowel prolongation, then sentence reading and finally passage reading; another one-third started with sentence reading, then passage reading and finally sustained vowel prolongation; and the remaining started with passage reading, then sustained vowel prolongation and finally sentence reading.

Voice samples were presented through a computer with an audio interface (M-Audio; Irwindale, CA, USA) and Sennheiser HD 590 headphones (Sennheiser; Wedemark, Hanover, Germany). Listeners were given a maximum of two hours to complete the ratings and were encouraged to rate at their own pace and to take regular breaks (a one minute break per 30 stimuli) to minimize any fatigue effect. They were allowed to adjust the voice loudness to a comfortable level throughout the rating. Each voice could be played for a maximum of three times according to their needs, but could not be replayed once other voice samples had been

played. The background noise intensity was checked using a sound level meter (210 Sound Level Meter by Quest Technologies), with a range from 54 to 68 dB (mean = 56.2 dB) detected in the surrounding environments.

Descriptive Statistics

Central tendency (group mean), range and dispersion (standard deviation) of the ratings by the two groups of listener on the three vocal parameters, the three speech contexts and the four severity levels were calculated.

Statistical Analysis

All Statistical analyses were conducted by a computer program named *Statistical Package for the Social Sciences* – IBM SPSS for Mac, Version 22.0.0 (IBM, Chicago, IL). Intraclass correlation coefficient (ICC) was used to estimate intra- and inter-rater reliability. The ICC model used to calculate intra-rater reliability in this study was based on the assumption that listeners were the only listeners of interest, and each of them rated all voices, while that in inter-rater reliability was based on the assumption that the listeners represented a random set of listeners from a larger population, and each of them gave ratings for all voice samples (Shrout, & Fleiss, 1979). Intra-rater reliability was computed using consistency-of-agreement ICC for a two-way mixed-effects model, i.e. ICC (3,1), while inter-rater reliability was computed using absolute-agreement ICC for a two-way random-effects model, i.e. ICC (2,1), for each group of listeners. The ICC values of professional and naïve listeners on the three speech contexts, the three vocal parameters and the four severity levels were then compared. An ICC greater than 0.75 indicates a satisfactory reliability (Portney, & Watkins, 2000).

Results

Descriptive Statistics

Table 2 presents the means, standard deviations and ranges of ratings for speech contexts on vocal parameters across grades of voice disorders by the two groups of listeners.

Table 2. Mean Ratings made by Professional and Naïve Listeners.

Professional Listeners										
Speech Context		**Sustained /a/**			**Sentence Reading**			**Passage Reading**		
Severity	**Parameter**	**M**	**(SD)**	**RG**	**M**	**(SD)**	**RG**	**M**	**(SD)**	**RG**
Normal	**R**	1.38	(1.40)	0 - 6	0.92	(1.21)	0 - 5	0.688	(0.985)	0 - 4
	B	1.09	(1.34)	0 - 6	0.327	(0.585)	0 - 3	0.275	(0.571)	0 - 2
	OS	1.47	(1.41)	0 - 7	0.625	(0.777)	0 - 4	0.575	(0.872)	0 - 3
Mild	**R**	2.91	(1.85)	0 - 8	3.19	(1.84)	0 - 7	2.29	(1.83)	0 - 9
	B	3	(1.80)	0 - 8	3.06	(1.96)	0 - 8	2.19	(1.73)	0 - 8
	OS	2.96	(1.79)	0 - 8	3.86	(1.88)	1 - 10	2.53	(2.05)	0 - 10
Moderate	**R**	4.32	(2.00)	0 - 8	3.64	(2.01)	0 - 9	3.7	(2.18)	0 - 10
	B	4.71	(2.39)	0 - 10	3.66	(2.18)	0 - 9	3.68	(2.71)	0 - 10
	OS	5.26	(2.14)	1 - 10	4.51	(2.21)	1 - 9	4.2	(2.49)	0 - 10
Severe	**R**	4.7	(2.14)	1 - 9	5.2	(1.96)	2 - 10	6.07	(1.66)	2 - 10
	B	4.66	(2.60)	0 - 10	3.92	(2.55)	0 - 10	4.82	(2.28)	1 - 10
	OS	5.43	(2.27)	2 - 10	4.93	(2.27)	1 - 10	6.1	(1.73)	2 - 10
Naïve Listeners										
Normal	**R**	1.80	(1.75)	0 - 9	0.994	(1.22)	0 - 8	1.35	(1.62)	0 - 8
	B	1.51	(1.66)	0 - 8	0.997	(1.23)	0 - 7	1.39	(1.70)	0 - 8
	OS	1.56	(1.58)	0 - 8	0.918	(1.12)	0 - 7	1.3	(1.57)	0 - 8
Mild	**R**	3.21	(1.97)	0 - 8	3.32	(1.95)	1 - 7	2.93	(1.96)	0 - 7
	B	3.05	(2.18)	0 - 8	2.78	(2.12)	0 - 7	2.85	(2.21)	0 - 7
	OS	3.07	(1.88)	0 - 8	2.95	(1.92)	0 - 7	2.98	(2.05)	0 - 8
Moderate	**R**	4.57	(1.91)	0 - 9	3.76	(2.19)	0 - 10	4.55	(2.26)	0 - 10
	B	4.21	(2.19)	0 - 9	3.14	(2.27)	0 - 10	3.96	(2.54)	0 - 10
	OS	1.89	(4.43)	0 - 9	3.36	(2.14)	0 - 10	4.36	(2.30)	0 - 10
Severe	**R**	4.88	(2.26)	0 - 10	4.87	(2.18)	0 - 10	4.55	(2.26)	0 - 10
	B	4.43	(2.56)	0 - 10	3.58	(2.01)	0 - 10	5.47	(2.42)	0 - 10
	OS	4.7	(2.26)	0 - 10	4.4	(2.06)	0 - 10	6.51	(1.94)	2 - 10

Note: M: mean; SD: standard deviation; RG: range; R: roughness; B: breathiness; OS: overall severity

Intra-rater Reliability by Professional and Naïve Listeners

Table 3 lists the intra-rater reliability obtained by professional and naïve listeners. ICC (3,1) was used. In general, the mean ICC of professional listeners was 0.831 (ranged from 0.727 to 0.909), while the mean of naïve listeners was 0.746 (ranged from 0.504 to 0.866).

Table 4 shows the mean intra-rater reliability of the three types of voice samples on the three vocal parameters by the two groups of listener. In professional listeners, ICC ranged from 0.466 to 0.968. The means were 0.780 for sustained /a/, 0.810 for sentence reading and 0.874 for passage reading. The means were 0.787 for roughness, 0.831 for breathiness and 0.847 for overall severity. In naïve listeners, ICC ranged from -0.119 to 0.974. The means were 0.655 for sustained /a/, 0.717 for sentence reading and 0.827 for passage reading. The means were 0.735 for roughness, 0.724 for breathiness and 0.750 for overall severity.

The mean intra-rater reliabilities of professional and naïve listeners across the four grades of severity under the three speech contexts and the three vocal parameters were not calculated due to small number of data points in each calculation. However, the mean intra-rater reliabilities of the two groups across severity grades regardless of the vocal parameters and the speech contexts were found. Table 5 lists the mean intra-rater reliabilities of the two groups of listeners across severity grades.

Figure 1 shows the intra-rater reliability of the three types of voice samples by professional and naïve listeners using a boxplot.

Table 3. Intra-rater Reliability of Each Professional and Naïve Listener.

Professional Listeners	ICC	Naïve Listeners	ICC
P1	.800	N1	.754
P2	.773	N2	.765
P3	.898	N3	.796
P4	.909	N4	.718
P5	.814	N5	.504
P6	.885	N6	.866
P7	.769	N7	.822
P8	.897	N8	.842
P9	.727	N9	.748
P10	.838	N10	.763
		N11	.746
		N12	.720
		N13	.738
		N14	.782
		N15	.764
		N16	.677
		N17	.779
		N18	.754
		N19	.579
		N20	.786
Mean = 0.831		Mean = 0.746	
Range = 0.727 – 0.909		Range = 0.504 – 0.866	

Note: ICC: intraclass correlation; numbers underlined represent ICC lower than 0.75

Table 4. Intra-rater Reliability of the Three Speech Types across the Three Vocal Parameters by Professional and Naïve Listeners.

		Professional Listeners			
	Speech Context	**Sustained /a/**	**Sentence Reading**	**Passage Reading**	**Mean**
Vocal Parameter	**Roughness**				
	Mean	0.707	0.791	0.862	0.787
	(Range)	(0.466 – 0.886)	(0.607 – 0.949)	(0.704 – 0.964)	
	Breathiness				
	Mean	0.825	0.809	0.860	0.831
	(Range)	(0.726 – 0.937)	(0.664 – 0.898)	(0.536 – 0.950)	
	OS				
	Mean	0.809	0.831	0.900	0.847
	(Range)	(0.640 – 0.896)	0.718 – 0.964	(0.765 – 0.968)	
	Mean	0.780	0.810	0.874	Overall mean ICC = 0.831

Table 4. Intra-rater Reliability of the Three Speech Types across the Three Vocal Parameters by Professional and Naïve Listeners (Continue).

		Naïve Listeners			
	Speech Context	**Sustained /a/**	**Sentence Reading**	**Passage Reading**	**Mean**
Vocal Parameter	**Roughness**				
	Mean (Range)	0.647 (0.116 – 0.836)	0.730 (0.303 – 0.904)	0.819 (0.545 – 0.959)	0.735
	Breathiness				
	Mean (Range)	0.627 (-0.119 – 0.835)	0.656 (0.380 – 0.840)	0.803 (0.566 – 0.948)	0.724
	OS				
	Mean (Range)	0.638 (-0.095 – 0.876)	0.713 (0.272 – 0.869)	0.849 (0.587 – 0.974)	0.750
	Mean	0.655	0.717	0.827	Overall mean ICC = 0.746

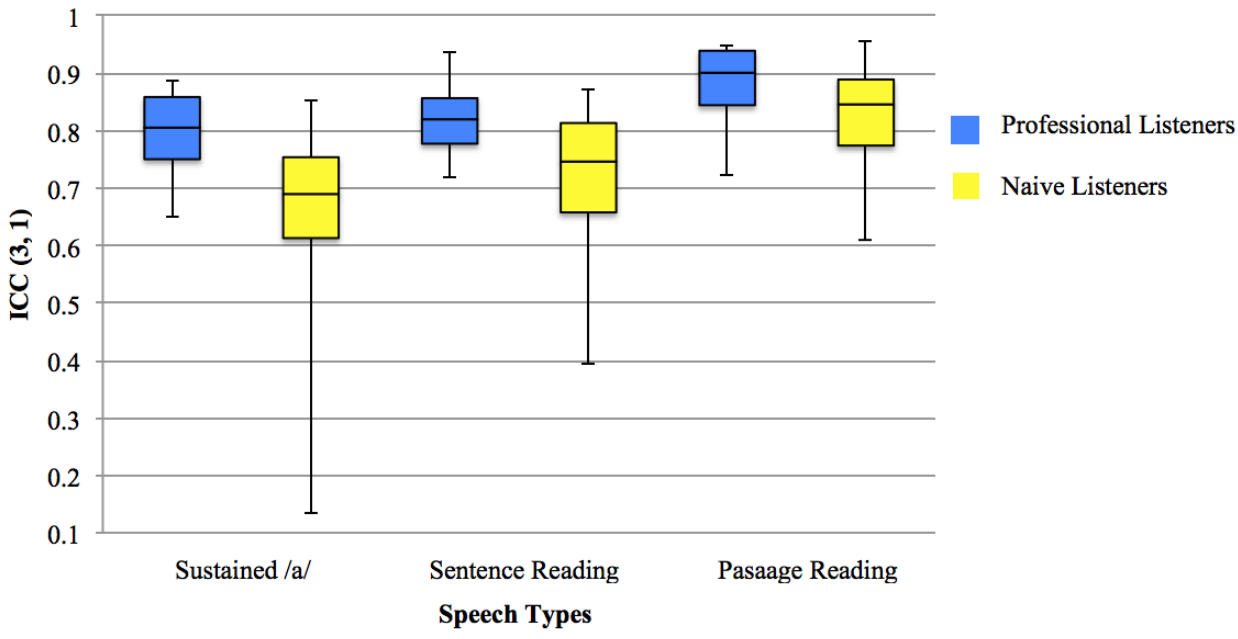
Note: ICC: intraclass correlation; OS: overall severity

Table 5. Intra-rater Reliability of the Voice Samples across Four Severity Grades.

Severity Grades	**Normal**	**Mild**	**Moderate**	**Severe**
Professional Listeners	0.567	0.673	0.665	0.602
Naïve Listeners	0.287	0.516	0.612	0.616

Note: ICC: intraclass correlation

Figure 1. Intra-rater Reliability of Three Speech Types by Professional and Naïve Listeners.



Inter-rater Reliability by Professional and Naïve Listeners

Table 6 lists the inter-rater reliability on the three types of voice samples across the three vocal parameters rated by the two groups of listeners. ICC (2, 1) was used. None of the vocal parameters reached an inter-rater reliability above 0.75 in professional or naïve listeners. In professional listeners, ICC ranged from 0.505 to 0.725, with an overall mean of 0.620. In terms of speech contexts, the means were 0.553 for sustained /a/, 0.608 for sentence reading and 0.686 for passage reading. In terms of vocal parameters, the means were 0.606 for roughness, 0.621 for breathiness and 0.628 for overall severity. In naïve listeners, ICC ranged from 0.310 to 0.593, with an overall mean of 0.460. In terms of speech contexts, the means were 0.387 for sustained /a/, 0.419 for sentence reading and 0.530 for passage reading. In terms of vocal parameters, the means were 0.500 for roughness, 0.372 for breathiness and 0.508 for overall severity.

Table 7 lists the inter-rater reliability of the three types of voice on the three vocal parameters across the four severity levels by professional and naïve listeners. In professional listeners, ICC was 0.321 for normal voices, 0.303, 0.359 and 0.247 for mild, moderate and severe dysphonic voices respectively. In naïve listeners, ICC was 0.121 for normal voices, and 0.137, 0.232 and 0.274 for mild, moderate and severe dysphonic voices respectively.

Table 6. Inter-rater Reliability of the Three Types of Voice Sample Across Three Vocal Parameters by Professional and Naïve Listeners.

		Professional Listeners			
	Speech Context	**Sustained /a/**	**Sentence Reading**	**Passage Reading**	**Mean**
Vocal Parameter	**Roughness**	0.505	0.572	0.725	0.606
	Breathiness	0.597	0.605	0.653	0.621
	OS	0.558	0.646	0.680	0.628
	Mean	0.555	0.609	0.684	Overall ICC = 0.620
		Naïve Listeners			
Vocal Parameter	**Roughness**	0.412	0.481	0.579	0.500
	Breathiness	0.336	0.310	0.447	0.372
	OS	0.428	0.456	0.593	0.508
	Mean	0.387	0.419	0.530	Overall ICC = 0.460

Note: ICC: intraclass correlation; OS: overall severity

Table 7. Inter-rater Reliability of the Three Types of Voice Sample Across Three Vocal Parameters on Four Severity Levels by Professional and Naïve Listeners.

		Speech Context	**Sustained /a/**	**Sentence Reading**	**Passage Reading**	**Mean**
	Severity level	**Vocal Parameter**				
Professional Listeners	**Normal**	**Roughness**	0.280	0.049	0.236	
		Breathiness	0.522	0.083	0.108	0.321
		OS	0.417	0.058	0.207	
	Dysphonic	**Roughness**	0.307	0.390	0.608	
		Breathiness	0.355	0.320	0.459	-
		Overall Severity	0.287	0.428	0.490	
	A. Mild	**Roughness**	0.267	0.224	0.396	
		Breathiness	0.325	0.336	0.308	0.303
		OS	0.266	0.312	0.261	
	B. Moderate	**Roughness**	0.186	0.214	0.472	
		Breathiness	0.291	0.455	0.495	0.359
		OS	0.168	0.378	0.389	
	C. Severe	**Roughness**	0.216	0.453	0.219	
		Breathiness	0.293	0.131	0.146	0.247
		OS	0.088	0.327	0.095	
	Overall ICC = 0.620					

Table 7. Inter-rater Reliability of the Three Types of Voice Sample Across Three Vocal Parameters on Four Severity Levels by Professional and Naïve Listeners (Continue).

Naïve Listeners	**Normal**	**Roughness**	0.183	0.007	0.120	0.121
		Breathiness	0.079	0.028	0.123	
		OS	0.129	0.018	0.116	
	Dysphonic	**Roughness**	0.240	0.220	0.455	-
		Breathiness	0.162	0.122	0.342	
		OS	0.232	0.230	0.487	
	A. Mild	**Roughness**	0.188	0.097	0.092	0.137
		Breathiness	0.112	0.119	0.223	
		OS	0.173	0.124	0.208	
	B. Moderate	**Roughness**	0.078	0.235	0.314	0.232
		Breathiness	0.017	0.165	0.295	
		OS	0.050	0.235	0.393	
	C. Severe	**Roughness**	0.198	0.148	0.216	0.274
		Breathiness	0.219	0.165	0.165	
		OS	0.212	0.181	0.201	
	Overall ICC = 0.460					

Note: ICC: intraclass correlation; OS: overall severity

Discussion

The present study set out to achieve two objectives. The first objective was to compare the reliability between professional listeners and naïve listeners on perceptual rating of children voice samples. The second was to find out whether the speech contexts would influence listeners' reliability on perceptual rating of children voice samples.

Effects of Listeners' Background on Rater Reliability

The first objective was to compare rater reliability between two groups of listeners (i.e., professional and naïve listeners) on perceptual rating of children voice samples. The comparison of the two groups of listener on rater reliability is illustrated in the following.

Intra-rater reliability. The proportion of professional listeners who achieved satisfactory level of intra-rater reliability was greater than that of naïve listeners (90% in

professional listeners; 60% in naïve listeners, see table 3). The mean intra-rater reliability was relatively higher in professional listeners than that in naïve listeners. ICC coefficients with a level of 0.75 or above suggest a satisfactory level of reliability (Portney, & Watkins, 2000). The mean intra-rater reliability of professional listeners was greater than 0.75 (mean ICC = 0.831, see table 3), implying a satisfactory level of intra-rater reliability. The levels attained from sustained /a/ prolongation, sentence reading and passage reading were all satisfactory (ICC > 0.75, see table 4). However, the mean intra-rater reliability of naïve listeners was below 0.75 (mean ICC = 0.746, see table 3), implying a moderate level of intra-rater reliability. The levels achieved from sustained /a/ and sentence reading were moderate (ICC in sustained /a/ = 0.655; ICC in sentence reading = 0.717, see table 4), and the level obtained from passage reading was satisfactory (ICC = 0.827, see table 4).

The satisfactory intra-rater reliability attained by professional listeners, but moderate level of intra-rater reliability attained by naïve listeners could be explained by the followings. Most of the professional listeners received previous trainings on perceptual voice evaluation and pediatric voice disorders. Therefore, they might possess a more stable internal standard on perceptual evaluation of voice and a better perceptual sensitivity to the vocal parameters than naïve listeners (Law et al., 2012). In addition, a majority of professional listeners (6 out of 10 listeners) had 4 years of experience or more. This might suggest that they were more experienced in voice evaluation and management and thus, more experienced in rating children's voice. However, naïve listeners did not receive any training on perceptual voice evaluation in advance. Therefore, it might be possible that their internal standards changed throughout the rating procedures, leading to lower intra-rater reliability.

Inter-rater reliability. The inter-rater reliability of professional listeners (mean ICC = 0.620, see table 6) was relatively higher than that of naïve listeners (mean ICC = 0.460, see table 6) in general. However, both listener groups gave inter-rater reliability of lower than

0.75, i.e., moderate inter-rater reliability by professional listeners, while poor inter-rater reliability by naïve listeners.

Higher inter-rater reliability obtained by professional listeners could be explained in terms of listeners' internal standards. All professional listeners received trainings on perceptual voice evaluation and management, and on pediatric voice disorders beforehand. Therefore, their internal standards on voice rating would be better established and calibrated than that in naïve listeners, resulting in higher inter-rater reliability. On the other hand, the relatively lower inter-rater reliability in naïve listeners might be due to a lack of training on voice evaluation and a lack of exposure to voice problems in children.

Effect of Types of Voice Samples on Rater Reliability

The second objective was to find out whether speech contexts would affect listeners' reliability on perceptual rating of children voice samples. The following illustrates and justifies the influence of speech contexts on raters' reliability.

The results of the current study supported that there were differences in rater reliability across the types of voice samples. ICC indicated that the intra-rater reliability was the highest for passage reading (ICC by professional listeners = 0.874; ICC by naïve listeners = 0.827, see table 4), followed by sentence reading (ICC by professional listeners = 0.810; ICC by naïve listeners = 0.717, see table 4) and then sustained /a/ prolongation (ICC by professional listeners = 0.780; ICC by naïve listeners = 0.655, see table 4) for both groups of listener. Similarly, the above pattern was also found in inter-rater reliability. The ICC values obtained from passage reading was the highest (ICC by professional listeners = 0.684; ICC by naïve listeners = 0.530, see table 6), followed by sentence reading (ICC by professional listeners = 0.609; ICC by naïve listeners = 0.419, see table 6) and then sustained /a/ prolongation (ICC by professional listeners = 0.555; ICC by naïve listeners = 0.387, see table

6). It could be concluded that regardless of the listeners' experience on perceptual evaluation and management of voice disorders, connected speech samples, particularly passage reading, were more reliably rated than sustained /a/. This might be contributed to the fact that the production of sustained /a/ does not possess as great demands on the coordination of vocal folds, laryngeal and supralaryngeal muscles as connected speech, it therefore fails to fully reflect the deviant aspects of voice (Poletto et al., 2004).

For connected speech, passage reading gave higher intra- and inter-rater reliability than sentence reading. It might also be due to greater demands placed on the coordination of laryngeal muscles in passage reading than in sentence reading that results in more disclosure on deviant voice quality. As a result, more information regarding the children's voice quality can be revealed upon passage reading (Poletto et al., 2004; Law et al., 2012), leading to increased rater reliability in voice evaluation of passage reading.

It is worth noting that the inter-rater reliability achieved by professional listeners in this study (ICC on sustained /a/ = 0.555; ICC on passage reading = 0.684, see table 6) was lower than that in the recent adult study (ICC on sustained /a/ = 0.583; ICC on passage reading = 0.708) (Law et al., 2012). This might be because studies focusing on evaluations of pediatric voice disorders are limited in the literature, the internal standards among professional listeners on rating pediatric voices might be more difficult to be calibrated than that on rating adult voices. Therefore, a lower inter-rater reliability by professional listeners was achieved in this study.

Rater Reliability on Different Vocal Parameters

Differences were found in rater reliability when rating the vocal parameters of roughness, breathiness and overall severity. Results revealed that intra- and inter-rater reliability on the parameter of overall severity (Intra-rater: ICC by professional listeners = 0.847, ICC by naïve

listeners = 0.750, see table 4; Inter-rater: ICC by professional listeners = 0.628, ICC by naïve listeners = 0.508, see table 6) was found to be the greatest among the three vocal parameters by both groups of listener. This agreed with the conclusions drawn in the studies by Revis, Giovanni, Wuyts, and Triglia (1999), and Munoz et al. (2002), which focused on adults' voices.

Interestingly, it was noted that for professional listeners, higher intra- and inter-rater reliability were concluded on breathiness (Intra-rater: ICC = 0.831, see table 4; Inter-rater: ICC = 0.621, see table 6) than on roughness (Intra-rater: ICC = 0.787, see table 4; Inter-rater: ICC = 0.606, see table 6), while for naïve listeners, higher intra- and inter-rater reliability were found in roughness (Intra-rater: ICC = 0.735, see table 4; Inter-rater: ICC = 0.500, see table 6) than in breathiness (Intra-rater: ICC = 0.724, see table 4; Inter-rater: ICC = 0.372, see table 6). The reasons for these should be further examined.

Rater Reliability under Different Severity Levels

For professional listeners, higher intra-and inter-rater reliability was attained for mild-to-moderate grade of dysphonia (Intra-rater: ICC for mild = 0.673, ICC for moderate = 0.665, see table 5; Inter-rater: ICC for mild = 0.303, ICC for moderate = 0.359, see table 7), but lower in rating normal speakers and severe grade of dysphonic speakers (Intra-rater: ICC for normal = 0.567, ICC for severe = 0.602, see table 5; Inter-rater: ICC for normal = 0.321, ICC for severe = 0.247, see table 7). Intra-rater reliability was the lowest when rating the voices of normal speakers, while inter-rater reliability was the lowest for severely dysphonic speakers. For naïve listeners, the highest intra-and inter-rater reliability was achieved when rating severe grade of dysphonic speakers (Intra-rater: ICC = 0.616, see table 5; Inter-rater: ICC = 0.274, see table 7) and the lowest when rating normal speakers (Intra-rater: ICC =

0.287, see table 5; Inter-rater: ICC = 0.121, see table 7). The conclusions for professional and naïve listeners were conflicting.

The results shown in professional listeners were contradictory to the previous study by Law et al. (2012) on adults' voices, which concluded that the mean ICC on inter-rater reliability of severely dysphonic speakers was the highest and that of normal speakers was the lowest, while the results shown in naïve listeners agreed with that. The present results also contradicted to the conclusions drawn by Yu, Revis, Wuyts, Zanaret, and Giovanni (2002) on adults' voices, which claimed that normal voices give the highest reliability due to amble experiences encountered in everyday life situations on normal voice quality by all listeners.

The contradictory conclusions might be attributed to the fact that the internal standard of voice qualities and severities in children voices are more varied than that in adult voices. The results urge the need for more studies in children voice.

Clinical Implications

The current study provided clinical evidences on which type of voice sample to be used in perceptual voice evaluation in children. Based on the results obtained in this study, passage reading is recommended for perceptual ratings of children voice in clinical practices. It is because the mean intra-rater reliability of professional listeners on passage reading was the highest among the three speech contexts. Higher intra-rater reliability can increase the validity of the ratings on voice and facilitate documentation of the progress along the course of voice management. Another reason is that passage reading yielded the highest inter-rater reliability by professional listeners in this study. Higher inter-rater reliability allows better handover of children's progress on vocal use, in case there is a change in clinician responsible for the children's voice management. In addition, passage reading can better

reveal the natural voice used in daily situations. Therefore, passage reading is recommended to elicit children's voice sample for clinical purposes.

However, the passage may often be quite long in length and may contain words that are unfamiliar for young children. It is suggested to use a passage with around five utterances, which is similar to that used in this study, and with commonly used words to elicit connected speech in children. If young children fail to read the passage, a sentence with 5 syllables will then be recommended. Though sentence reading yielded lower intra-rater reliability by professional listeners than passage reading in this study, it is a type of connected speech that can reveal pathogenic voice characteristics more thoroughly than sustained /a/ prolongation, and it is shorter in length, so that it can be produced by young children with greater ease.

Limitations and Directions for Future Research

The following issues may be interesting to be explored in future researches. First of all, conversational speech samples were not used in the current study due to difficulties in eliciting and standardizing children's spontaneous speech. However, it may be interesting to find out whether conversational speech and passage reading give different rater reliability, and which yields higher reliability. It is because conversational speech can better reflect natural voice characteristics in daily situations, if the reliabilities of conversational speech and passage reading are compared, it may give an insight to which speech context yields the best reliability in perceptual voice evaluation in children. In addition, the number of years of experience in pediatric voice caseloads of the professional listeners varied in this study. It may be interesting to investigate whether the years of experience would affect rater reliability on pediatric voice rating.

Moreover, this study suggested a need to further investigate 1) the effect of children dysphonic severity of children speakers on perceptual rating of voice samples, and 2) the

reasons for higher rater reliability on breathiness than on roughness obtained by professional listeners, but vice versa by naïve listeners. The present study also provided insights on the importance of trainings on voice ratings to achieve satisfactory intra- and inter-rater reliability among listeners. In addition, the provision of anchors before ratings may also help improve rater reliability by facilitating the calibration of internal standards in listeners on the severity levels of vocal parameters.

Conclusions

To conclude, this study showed that professional listeners in general gave higher intra- and inter-rater reliability on perceptual rating of children voice samples. Passage reading yielded the highest intra- and inter- rater reliability in both groups of listeners. Conclusion could not be drawn on the effect of dysphonic severity levels on rater reliability. These may indicate that trainings are necessary prior to perceptual evaluation of children's voice and using passages to obtain voice samples from children provides the most reliable ratings. Further investigations on the reliability obtained from conversational speech samples and the effect of dysphonic severity on rater reliability of children's voices are required. Provision of trainings or anchors prior to perceptual ratings of children voice may improve rater reliability.

Acknowledgements

I would like to express my sincere gratitude to my dissertation supervisor, Dr. Estella Ma, for her valuable comments on my work, support and guidance throughout the development of this study. I would like to show my gratefulness to all listeners for their kind participations and supports. At last, I would like to give thanks to my family and friends for their support, encouragement and love.

References

Baker, S., Weinrich, B., Bevington, M., Schroth, K., and Schroeder, E. (2008). The effect of task type on fundamental frequency in children. *International Journal of Pediatric Otorhinolaryngology, 72(6),* 885 – 889.

Colton, R. H., Casper, J. K., Leonard, R. (2006). *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment (3rd ed.).* Baltimore: Lippincott Williams & Wilkins.

de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research, 37(5),* 985 – 1000.

Eadie, T. L., Kapsner, M., Rosenzweig, J., Waugh, P., Hillel, A., and Merati, A. (2010). The role of experiences on judgments of dysphonia. *Journal of Voice, 24(5),* 564-573.

Freitas, S. V., Pestana, P. M., Almeida, V., and Ferreira, A. (2013). Audio-perceptual evaluation of Portuguese Voice Disorders – An inter- and intrajudge reliability study. *Journal of Voice, 28(2),* 210 – 215.

Ghio, A., Revis, J., Merienne, S., and Giovanni, A. (2013). Top-down mechanisms in dysphonia perception: The need for blind tests. *Journal of Voice, 27(4),* 481 – 485.

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., and Hoffman, H. T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21(5),* 576 – 590.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5,* 7-23.

Kreiman, J., Gerratt, B. R., and Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *Journal of Acoustics Society of America, 122(4),* 2354-2364.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., and Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21 – 40.

Law, T., Kim, J. H., Lee, K. Y., Tang, E. C., Lam, J. H., van Hasselt, A. C., and Tong, M. C. (2012). Comparison of rater's reliability on perceptual evaluation of different types of voice sample. *Journal of Voice, 26(5),* 666.e13 – 666.e21.

Munoz, J., Mendoza, E., Fresneda, M. D., Carballo, G. (2002). Perceptual analysis in different voice samples: Agreement and reliability. *Perceptual and Motor Skills, 94 (3c),* 1187 – 1195.

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatrica et Logopaedica, 61(1),* 49-56.

Poletto, C. J., Verdun, L. P., Strominger, R., and Ludlow, C. L. (2004). Correspondence between laryngeal vocal fold movement and muscle activity during speech and nonspeech gestures. *Journal of Applied Physiology, 97,* 858 – 866.

Portney, L., and Watkins, M. P. (2000). *Foundations of Clinical Research: Applications to Practice. 2rd ed.* New Jersey: Prentice Hall Health.

Revis, J., Giovanni, A., Wuyts, F., and Triglia, J. (1999). Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopaedica, 51(3),* 108 – 116.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86 (2),* 420 – 428.

Swerts, M., and Veldhuis, R. (2001). The effect of speech melody on voice quality. *Speech Communication, 33,* 297 – 303.

Yiu, E. M. L., Murdoch, B., Hird, K., Lau, P., and Ho, E. M. (2008). Cultural and language differences in voice quality perception: A preliminary investigation using synthesized signals. *Folia Phoniatrica et Logopaedica, 60,* 107 – 119.

Yiu, E. M. L., and Ng, C. Y. (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics & Phonetics, 18(3),* 211 – 229.

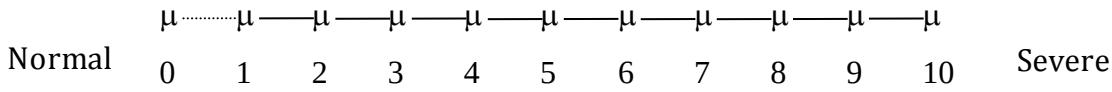
Yu, P., Revis, J., Wuyts, F. L., Zanaret, M., and Giovanni, A. (2002). Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatrica et Logopaedica, 54,* 271 – 281.

Appendix A

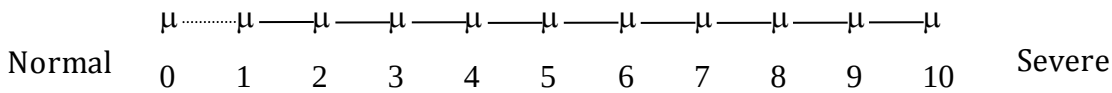
The 11-point Equal Appearing Interval Rating Scale

Sample

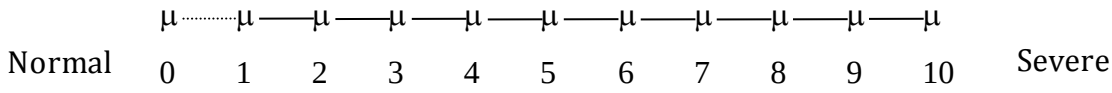
Roughness



Breathiness



Overall Severity



Appendix B

Sentence reading and passage reading for voice sampling

Sentence reading:

爸爸打哥哥

Passage reading:

〈說話不簡單〉

*****上課了，山羊老師要教大家說話。小牛和小馬聽了，都覺得好笑，心裏想：誰不會說話呢？*****

山羊老師請同學們出來練習說話。小牛第一個舉手要說笑話。他越說越快，越說聲音越小，還沒說到一半就笑個不停。

小馬出來給大家講故事。同學們都專心聆聽，可是小馬前言不搭後語，大家越聽越不明白。

小牛和小馬終於知道，說話真不簡單，也要好好學習。

* This paragraph was extracted for perceptual voice rating by professional and naïve listeners.